# Nonparametric Forecasting of the Manufacturing Output Growth with Firm-level Survey Data

Gérard Biau, Olivier Biau, Laurent Rouvière

# Nonparametric Forecasting of the Manufacturing Output Growth with Firm-level Survey Data

Gérard BIAU [a], Olivier BIAU [b], and Laurent ROUVIÈRE [c]

[a] *Laboratoire de Statistique Théorique et Appliquée (LSTA)*
*Université Pierre et Marie Curie – Paris VI*
*Boîte 158, 175 rue du Chevaleret*
*75013 Paris, France*

gerard.biau@upmc.fr

[b] *INSEE, Direction des Etudes et Synthèses Economiques*
*Département de la Conjoncture, Division des Enquêtes de Conjoncture*
*15 boulevard Gabriel Péri, 92245 Malakoff Cedex, France*

olivier.biau@ensae.org

[c] *Institut de Recherche Mathématique de Rennes*
*UMR CNRS 6625, Laboratoire de Statistique*
*Université Rennes 2-Haute Bretagne, Campus Villejean*
*Place du Recteur Henri Le Moal, CS 24307, 35043 Rennes Cedex, France*

laurent.rouviere@univ-rennes2.fr

**Abstract**

**A large majority of summary indicators derived from the individual responses to qualitative Business Tendency Survey questions (which are mostly three-modality questions) result from standard aggregation and quantification methods. This is typically the case for the indicators called balances of opinion, which are the most currently used in short term analysis and considered by forecasters as explanatory variables in linear models. In the present paper, we discuss a new statistical approach to forecast the manufacturing growth from firm-survey responses. We base our predictions on a forecasting algorithm inspired by the random forest regression method, which is known to enjoy good prediction properties. Our algorithm exploits the heterogeneity of the survey responses, works fast, is robust to noise and allows the treatment of missing values. Starting from a real application on a French dataset related to the manufacturing sector, this procedure appears as a competitive method compared with traditional competing algorithms.**

# 1 Introduction

Due to their early release (by the end of the month in which they are conducted), Business Tendency Surveys (BTS) are widely used as potential indicators of the economic activity, ahead of the publication of data from quarterly national accounts. In particular, BTS results allow the elaboration of short-term forecasting models of the main aggregates of the national accounts on the basis of summary indicators derived from the surveyed responses.

Most BTS questions are qualitative and require either a positive response ("up" or "superior to average"), an intermediate one ("stable" or "close to average") or a negative one ("down" or "inferior to average"). A large majority of summary indicators derived from the individual responses to these questions result from standard quantification methods, mostly based on a combination of the percents of positive, stable and negative answers. This is typically the case with the so-called balance of opinion, which is the most currently used indicator for short-term analysis, and which is defined as the difference between the (generally weighted) proportion of positive responses with respect to the negative ones.

As such, these kinds of indicators encounter some criticism, essentially because they do not exploit the heterogeneity of the surveyed individual responses. In this respect, Mitchell, Smith, and Weale (2004, 2005) discuss alternative indicators of the economic activity, by relating firm categorical responses to official data via ordered discrete-choice models. Their applications to British and German survey data suggest that their indicators provide more accurate early estimates of manufacturing output growth than a set of classical aggregate indicators. However, on French data, Biau, Erkel-Rousse, and Ferrari (2006) find that the balances of opinion lead to better or, at least, as accurate short-term forecasts of the manufacturing production growth rate as the Mitchell, Smith, and Weale indicators.

In the present paper, we discuss a new statistical approach to forecast the manufacturing growth, with two important novelties. Firstly, we propose to exploit the heterogeneity of the firm-level survey responses by working out untreated data instead of balances of opinion. Secondly, we base our predictions on a forecasting algorithm inspired by the random forest regression method (Breiman, 2001a,b), which is known to be robust to noise and enjoy good prediction properties. Our algorithm exploits the heterogeneity of the survey responses, works fast, and allows the treatment of missing values.

The paper is organized as follows. In Section 2, we describe the dataset used in this study. Section 3 is devoted to the presentation of our forecasting algorithm. Finally, in Section 4, we briefly describe the INSEE (National Institute for Statistics and Economic Studies) traditional methodology and compare its performance with our model.

# 2 The data

Our application will be based on a French dataset related to the manufacturing sector. The quarterly manufacturing production growth rate is a quantitative data derived from the Quarterly National Accounts[1]. The entrepreneur individual qualitative responses are collected by the Business Survey Unit of the French Statistical Institute. Even if the French Industry survey is carried out on a monthly basis, we decided to use quarterly observations instead of monthly observations. This was motivated by the fact that the regular short-term forecasts of the economic activity performed by INSEE are precisely made on a quarterly basis. Our analysis covers the period ranging from the first quarter 1995 to the third quarter 2006. Moreover, we decided to test the forecasting performance of the methods on the type of data which are used in the operational conditions of the INSEE forecasting exercises. Therefore, we focused on the survey responses carried out in February, May, September and November[2].

The INSEE surveys deal with questions relating to production at the product level (not at the firm level). More precisely, each firm can declare up to four products[3] and answers questions regarding each of these products. In our analysis, we chose to retain only the biggest products (in terms of amount of sales). The total number of firms entering the survey during the considered period is 6,686[4]. On average, the number of responses during the period is equal to 17. In order to apply our methods, we selected firms whose number of responses was larger than the 3rd Quartile (Q3). Hence, we retained 1,760

[1]The empirical analysis was carried out in early january 2008. At that period, the last published release of the French quarterly accounts was the one presenting the detailled figures relating to the third quarter of 2007.

[2]The "Notes de conjoncture" are issued three times a year in March, June, and December. A more concise "Point de Conjoncture" updates the June Note in October. These publications present INSEE short term forecasts.

[3]1.4 product per firm is declared on average.

[4]Note that about 4,000 industrial entrepreneurs are interviewed during each survey. However, owing to economic developments (closure or restructuring of enterprises), the sample is updated periodically.

firms, and this gives on average 39 responses out of the 47 possible during the period (see Table 1 which presents a summary).

Table 1: Selection of firms.

| |
|---|
| BTS quarterly data from 1995-1 to 2006-3<br>(February, May, September, November). |
| Maximum responses in the period: 47.<br><br>Total number of firms: 6,686.<br>Average number of responses: 17.<br>Median: 12.<br>Q3 (3rd quartile): 26.<br><br>**Selection of 1,760 firms** whose number of responses is larger than 26.<br>Average of their response: 39.<br>Median of their response: 32.<br>Q3: 45. |

Let us consider a BTS, related to quarter $t$, in which $m = 1760$ manufacturing firms are asked whether their production has risen, remained unchanged or fallen. The responses are collected in a $m \times 2$ matrix denoted by $X_t$:

$$X_t = \begin{pmatrix} x_{1,1}^t & x_{1,2}^t \\ x_{2,1}^t & x_{2,2}^t \\ \vdots & \vdots \\ x_{i,1}^t & x_{i,2}^t \\ \vdots & \vdots \\ x_{m,1}^t & x_{m,2}^t \end{pmatrix}$$

where $x_{i,j}^t$ stands for the answer of the firm $i$ regarding the past production ($j = 1$) and the expected production ($j = 2$). As explained earlier, each $x_{i,j}^t$ can take four values:

$$x_{i,j}^t = \begin{cases} -1 & \text{for the answer “down”} \\ 0 & \text{for the answer “unchanged”} \\ 1 & \text{for the answer “up”} \\ NA & \text{when there is no response.} \end{cases}$$

With this notation, each observation $X_t$ consists of $2m$ features. Associated with $X_t$ is the manufacturing production quarterly growth rate observed at quarter $t$, denoted hereafter by $Y_t$. Thus, given a **new** BTS represented by a generic matrix $X = (x_{i,j})$, the statistical problem is to predict the associated manufacturing production quarterly growth rate $Y$ from the dataset $(X_1, Y_1), \ldots, (X_T, Y_T)$, where $T$ is the number of data items which are available to make the prediction. In our problem, $T = 47$.

Despite their qualitative nature, the surveys can be used to make quantitative short-term predictions of the macroeconomic magnitudes. This is a very useful exercise, as it can be carried out well before the national accounts figures become available. The results of the BTS are available about two months before the publication of the first estimates of the growth of Gross Domestic Product (GDP), that is at a particularly early point in time from the point of view of forecasters. We are now in a position to present our forecasting algorithm.

**Remark**. As pointed out by a referee, the explanatory information at hand is qualitative, and the codes -1 for down, +1 for up and 0 for same are, to a large extent, arbitrary. We realize that more involved codings may be more appropriate, depending on the forecast algorithm used. We believe however that such an analysis is beyond the scope of the paper.

# 3 The forecasting algorithm

## 3.1 Random forests

In the last years of his life, Breiman (2001a,b) promoted random forests for use in classification and regression. In one word, a random forest is a method which consists of many decision trees and outputs predictions which are obtained by aggregating over the tree set, typically using equal weights. Random forests are one of the most successful ensemble methods which exhibits performance on the level of boosting and support vector machines. The method is fast, robust to noise, does not overfit and offers possibilities for explanation and visualization of its input, such as variable selection. Moreover, as demonstrated below, it can easily be adapted to deal with missing data. Random forests have been shown to give excellent performance on a number of practical problems and are undoubtedly among the most accurate general-purpose regression methods available.

Algorithms for inducing a random forest were first developed by Breiman and Cutler, and "Random Forests" is their trademark. The web page

provides a collection of downloadable technical reports, and gives an overview of random forests as well as comments on the features of the method.

## 3.2 From trees to forests

Trees-based methods partition the feature space into a set of rectangles, and then fit a simple model (usually a constant) in each one. They are conceptually simple yet powerful. The tree regression algorithms are presented in detail in the monograph of Hastie, Tibshirani, and Friedman (2001). Let us briefly describe how to grow a binary regression tree using a dataset $(X_1, Y_1), \ldots, (X_T, Y_T)$. Recall that, in our context, each observation has $2m$ features (variables) and is of form

$$X = \begin{pmatrix} x_{1,1} & x_{1,2} \\ x_{2,1} & x_{2,2} \\ \vdots & \vdots \\ x_{i,1} & x_{i,2} \\ \vdots & \vdots \\ x_{m,1} & x_{m,2} \end{pmatrix}.$$

The algorithm CART automatically decides both splitting variables and split points. Suppose for example that we have a partition into $M$ regions, say $R_1, R_2, \ldots, R_M$, and we model the tree regressors as a constant $c_m$ in each region. Then the best $\hat{c}_m$ is just the average of the $Y_t$ falling in region $R_m$. Finding the best binary partition in terms of minimum sum of squares is generally computationally infeasible. Hence, it is usually done through the following heuristic. Starting with all observations, consider a splitting variable $x_{i,j}$ and split point $s$, and define the pair of half-planes

$$R_1\big[(i,j),s\big] = \{x_{i,j} \leq s\} \quad \text{and} \quad R_2[(i,j),s\big] = \{x_{i,j} > s\}.$$

Then we seek the splitting variable $(i,j)$ and split point $s$ which solve

$$\min_{(i,j),s} \Big[ \min_{c_1} \sum_{X_t \in R_1[(i,j),s]} (Y_t - c_1)^2 + \min_{c_2} \sum_{X_t \in R_2[(i,j),s]} (Y_t - c_2)^2 \Big].$$

For any choice $(i,j)$ and $s$, the inner minimization is solved by $\hat{c}_1$ (respectively $\hat{c}_2$) equal to the average of the $Y_t$ associated with the $X_t$ falling in $R_1$ (respectively $R_2$). For each splitting variable, the determination of the split point $s$ can be done very quickly. Therefore, by scanning through all

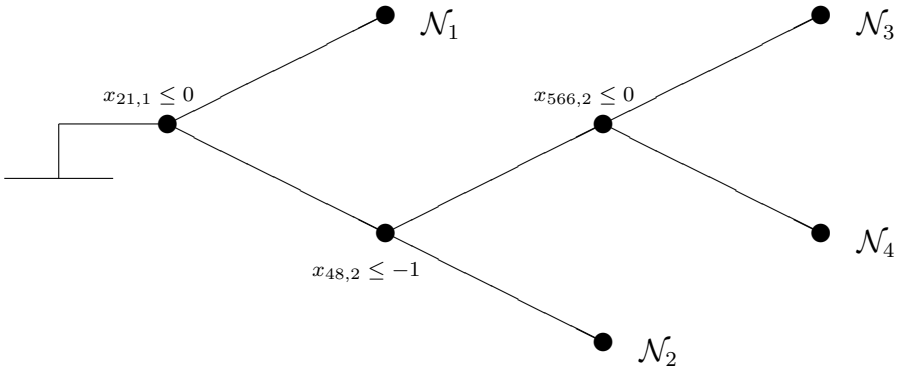

Figure 1: An example of binary tree.

the inputs, determination of the best pair $[(i, j), s]$ is feasible. Having found the best split, we partition the dataset into two resulting regions, we repeat the splitting process on each of the two regions, and so on. The process continues until each node (*i.e.*, a region) reaches a user-specified minimum node size $N_{min}$ and becomes a terminal node. In our problem, the terminal nodes, taken together, form a partition of $\mathbb{R}^{2m}$, and the tree regressor $h$ is then defined on each terminal region by the empirical mean

$$h(X) = \frac{1}{\text{Card}\left\{t : X_t \in \mathcal{N}(X)\right\}} \sum_{t:X_t \in \mathcal{N}(X)} Y_t,$$

where $\mathcal{N}(X)$ stands for the terminal node containing $X$.

The principle of random forests is to build multiple decision trees (often many hundreds) from different subsets of entities from the dataset and from different subsets of the features, to obtain substantial performance gains over single trees. Each decision tree is built from a bootstrapped sample of the full dataset (Efron and Tibshirani, 1993), and a random sample of the available variables is used for each node of each tree. Thus, instead of determining the optimal split on a given node by evaluating all possible splits on all variables, a subset of the variables, drawn at random, is used. Formally each tree is grown as follows:

1. Construct a bootstrap sample from $(X_1, Y_1), \ldots, (X_T, Y_T)$.
2. Choose $N_{min}$, the minimum node size.
3. Specify $p << 2m$ such that, at each node, $p$ variables only are selected at random out of the $2m$. The best splits (calculated with the CART

algorithm) on these $p$ variables for the bootstrap sample is used to split the node. Note that the value of $p$ is held constant during the growth of the forest.

For the free parameters $K$, $N_{min}$ and $p$, we used the default values $K = 500$, $N_{min} = 5$ and $p = \frac{2m}{3}$ of the random forest R-package[5].

Having built an ensemble of models, the final decision is the average value of the models. In other words, denoting by $h_1, \ldots, h_K$ the individual tree predictors, the final output is

$$h(X) = \frac{1}{K} \sum_{k=1}^{K} h_k(X). \tag{3.1}$$

We would like the reader to be aware that, although the mechanism of random forest algorithms appears simple, it is difficult to analyze and remains largely unknown. Some attempts to investigate the mathematical driving force behind consistency of random forests are by Breiman (2001a,b), Lin and Jeon (2006) (who establish a connection between random forests and adaptive nearest neighbor methods), and Biau, Devroye, and Lugosi (2007).

Nevertheless, random forests are known to enjoy exceptional prediction accuracy, and this accuracy is achieved for a wide range of settings of the tuning parameters. In addition, random forests possess a number of interesting features, including measures of proximities between the observations and measures of variable importance. In the next paragraph, we investigate how these features can be used to deal with the problem of missing values and variable selection.

## 3.3 Missing values and variable selection

The random forest predictor (3.1) does not support missing values in the $X_t$. As suggested by Breiman (2001b), missing values can be estimated by constructing proximities between the observations in the training sample. To this aim, after a tree is grown, we put all the data items $X_t$, $t = 1, \ldots, T$, down the tree. If $t$ and $t'$ are in the same terminal node, we increase the proximity between $X_t$ and $X_{t'}$ by one. To finish, we normalize by dividing by the number of trees. Thus, if $K$ stands for the number of tree predictors,

[5] http://lib.stat.cmu.edu/R/CRAN/src/contrib/Descriptions/randomForest.html

the proximity $P(X_t, X_{t'})$ between $X_t$ and $X_{t'}$ is defined by

$$P(X_t, X_{t'}) = \frac{1}{K}\sum_{k=1}^{K} \mathbf{1}_{\{X_t \in \mathcal{N}_k(X_{t'})\}} = \frac{1}{K}\sum_{k=1}^{K} \mathbf{1}_{\{X_{t'} \in \mathcal{N}_k(X_t)\}},$$

where $\mathcal{N}_k(X)$ is the terminal node of the tree $h_k$ which contains $X$.

Starting from Breiman's idea of proximity, we propose a new algorithm, called **RF1**, which allows the treatment of missing values. For notational convenience, $X$ will be denoted by $X_{T+1}$.

**RF1**

INPUT: $(X_1, Y_1), \ldots, (X_T, Y_T)$, $X_{T+1}$.

1. Consider any prediction $\tilde{Y}_{T+1}$ associated with $X_{T+1}$. Denote by $\mathcal{S}$ the augmented sample $(X_1, Y_1), \ldots, (X_T, Y_T), (X_{T+1}, \tilde{Y}_{T+1})$.

2. Fill in the missing values by the method of your choice. Denote by $\tilde{\mathcal{S}}$ the sample $(\tilde{X}_1, Y_1), \ldots, (\tilde{X}_T, Y_T), (\tilde{X}_{T+1}, \tilde{Y}_{T+1})$ without missing values.

3. Run the random forest algorithm on $\tilde{\mathcal{S}}$ and compute proximities.

4. Replace the missing values in the sample $\mathcal{S}$ by the average of the corresponding variables weighted by the proximities between the relevant cases and the non missing-value cases. More precisely, if $x_{i,j}^t = NA$, replace it by

$$\frac{1}{\sum_{\{t':t'\neq t, x_{i,j}^{t'} \neq NA\}} P(\tilde{X}_t, \tilde{X}_{t'})} \sum_{\{t':t'\neq t, x_{i,j}^{t'} \neq NA\}} P(\tilde{X}_t, \tilde{X}_{t'}) x_{i,j}^{t'}.$$

   Denote by $\tilde{\mathcal{S}} = (\tilde{X}_1, Y_1), \ldots, (\tilde{X}_T, Y_T), (\tilde{X}_{T+1}, \tilde{Y}_{T+1})$ the resulting sample.

5. Iterate $N$ times step 3. and step 4.

OUTPUT: the outcome predicted for $\tilde{X}_{T+1}$ by the random forest algorithm based on $(\tilde{X}_1, Y_1), \ldots, (\tilde{X}_T, Y_T)$.

Breiman argues that $N = 5$ iterations are generally enough. In our experiments, we chose for the initial $\tilde{Y}_{T+1}$ the (linear) prediction obtained by the traditional INSEE methodology, which will be described in Section 4.

Recall that each observation $X_t$ takes its values in a space of dimension $2m = 3,520$. However, it is well established that in high dimensional spaces, learning suffers from the curse of dimensionality (see for example Abraham, Biau, and Cadre, 2006). Thus, in practice, before applying any learning technique to model real data, a preliminary dimension reduction or model selection step is crucial for appropriate smoothing and circumvention of the dimensionality effect. In this respect, Breiman (2001b) suggests a measure, called variable importance, to discriminate between informative and noninformative variables. In the algorithm **RF2** below, we include this measure. The general idea is to run the random forest algorithm only on the most important variables.

**RF2**

INPUT: $(X_1, Y_1), \ldots, (X_T, Y_T)$, $X_{T+1}$.

1. Run the algorithm $RF1$ with input data $(X_1, Y_1), \ldots, (X_T, Y_T), X_{T+1}$ and compute the variable importance for each of the $2m$ variables.

2. Specify $p_{max} \leq 2m$ and for $t = 1, \ldots, T+1$, denote by $\bar{X}_t$ the vector composed of the $p_{max}$ most important variables of $X_t$.

OUTPUT: the outcome predicted by $RF1$ with input data $(\bar{X}_1, Y_1), \ldots, (\bar{X}_T, Y_T), \bar{X}_{T+1}$.

In our experiences, we observed that the choice $p_{max} = 700$ variables was enough. Thus, this dimension reduction step means that the algorithm automatically selects the 700 most representative entrepreneur answers out of the 3,520 possible ones.

# 4 Results and comparison with the INSEE methodology

Before presenting the practical results, we briefly describe the traditional INSEE methodology, which is based on linear models on the balances of

opinion. These models are the most currently used indicators for short-term analysis.

## 4.1 INSEE methodology

Balances of opinion are interesting indicators in many respects. Firstly, they are easy to implement. As univariate series, they are simple to read and to track over time, at the price of an acceptable loss of information with respect to the corresponding exhaustive three-dimensional statistics. Secondly, balances of opinion are subject to limited revisions across time. Finally, the main balances of opinion—notably those relating to activity—are highly correlated with the corresponding aggregates of interest, even though they are generally smoother (and therefore easier to read). This is typically the case, for instance, for the balances of opinion relating to past production derived from the INSEE Industry survey (see Figure 2). All these interesting properties explain why the balances of opinion are the main (if not the only) indicators used by short-term analysts as explanatory variables in a linear model. All in all, due to their good empirical properties, the balances of opinion prove to be very useful, as they are well adapted to the quick production and release conditions of BTS.

The most common methodology to predict the quarterly national accounts using business surveys, known as calibrations (see Raynaud and Scherrer, 1996, Buffeteau and Mora, 2000, Dubois and Michaux, 2006), consists in fitting a linear model between the balances of opinion $S_j^t$ (as before, $j = 1$ for the past production, and $j = 2$ for the expected production), and the dependent variable $Y_t$, which may typically be the manufacturing production growth. In mathematical terms,

$$Y_t = c + a_1 S_1^t + a_2 S_2^t + u_t,$$

where $u_t$ is some random noise.

The quality of this kind of model can be slightly improved by including the past values of $Y$ and by taking into account the variation of the balance of opinion. Nevertheless, in the present paper, we will focus on this simple model, whose validity and robustness have already been established, through the application of several specification tests using the estimated residuals, such as tests of stability of the coefficients (Chow test), tests of homoskedasticty (White test), or test of normality. We finally note that the calibration model uses the balances of opinion as computed and published by the INSEE.

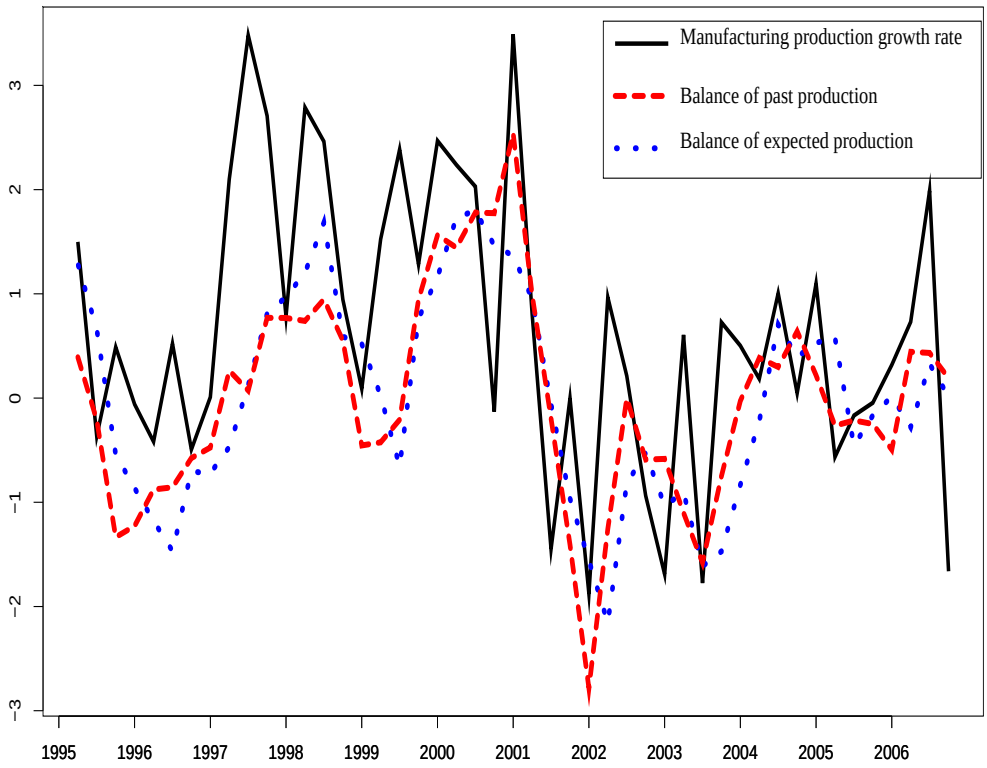


Figure 2: Balances of opinion relating to manufactured production together with the manufacturing production quarterly growth rate. (Note that the dataset has been centered and standardized).

These balances are based on the 4,000 firms data items, which are preprocessed to deal with missing values and seasonal adjustement. In the present study, the INSEE approach should be considered as a benchmark.

## 4.2 Results

The error rate for forecasting new observations is unknown. However, it can be estimated using a simple leave one out methodology. To this aim, we select one item $X_t$ together with its outcome $Y_t$ out of the 47 observations, and we consider it as new observation. Next, we determine the outcome $\hat{Y}_t$ using the procedure under study worked out with the 46 remaining data items, and we finally compare the estimated outcome with the true one. This process, repeated for each of the 47 observations, provides us with an estimate of the

mean square error rate, denoted hereafter by MSE:

$$\text{MSE} = \frac{1}{47}\sum_{t=1}^{47}(Y_t - \hat{Y}_t)^2.$$

We will use the following acronyms:

- LM refers to the linear model on the balances of opinion.
- RF1 and RF2 stand for the random forest-type algorithms described in Section 3.

The results obtained by the different procedures are presented in Table 2 and in Figure 3.

Table 2: Results of the different procedures.

| Method | MSE |
|---|---|
| LM | 1.27 |
| RF1 | 1.23 |
| RF2 | 1.18 |

Table 2 further emphasizes the good results achieved by the random forest algorithms. We note in particular the performance of RF2 which achieves, on average, the best MSE. The difference between RF1 and RF2 enlightens the importance of the variable selection step. We finally note that te RF2 algorithm works fast (using the R-package "RandomForest", our prediction take less than one minute) and is robust to the parameters (Bardaji, 2007). Our approach gives a new tool to the short term analysts, especially those of the INSEE, who can work on individual data.

# 5 Perspectives

To improve the results of the present study, we suggest two research directions. Firstly, it seems important to study the impact of putting weights on the entrepreneur responses: under the assumption that the firm size is correlated with the macro-economic production, an improvement in the relative performances of the random forest approach is possible. Secondly, one could

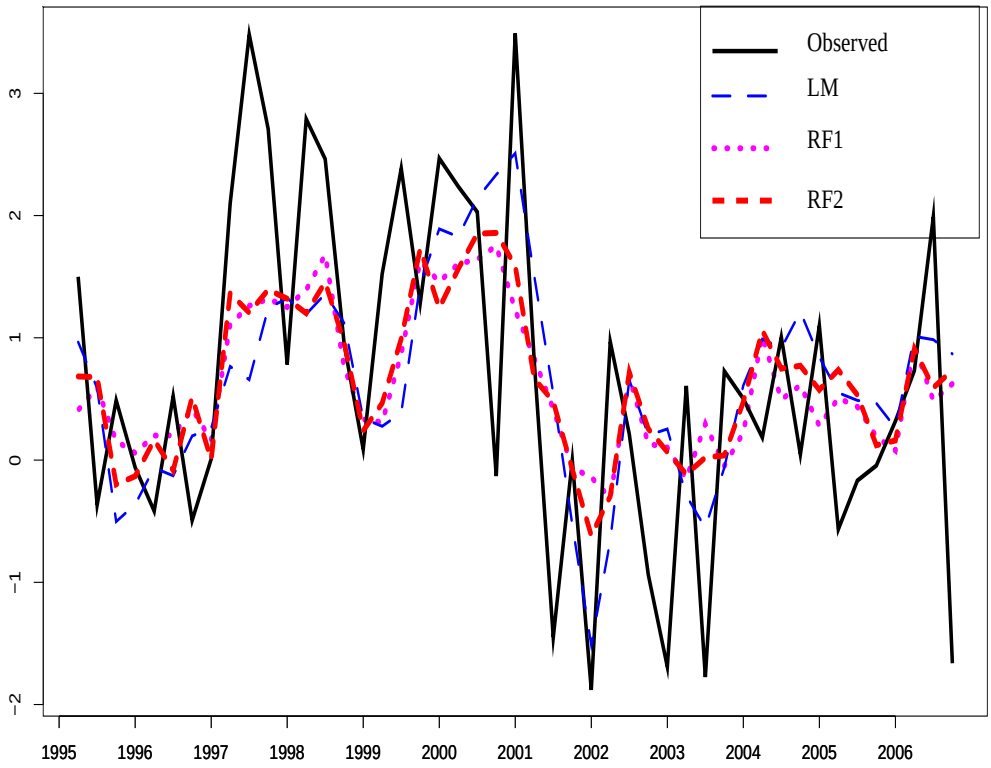


Figure 3: Manufacturing production quarterly growth rate and predictions obtained by the different methods.

use this new algorithm with other surveys (*e.g.* using retail trade survey to forecast household consumption) or mix the surveys (*e.g.* industry and services) to forecast the GDP. Finally, it would also be interesting to identify the 700 variables which are automatically selected by the algorithm RF2 (size, sector...). With this preliminary selection step, the calibration model using balances of opinion could also undoubtedly be improved.

**Acknowledgments.** The authors would like to thank Matthieu Cornec for his discussion of a preliminary version of the paper at the DEEE Workshop, INSEE, Paris, June 26, 2006, and Prof. Dr. Marco Lippi who served as a chairman at the 28th CIRET Conference in Rome, Italy, Septembre 2006. They are also indebted to José Bardaji for his comments and suggestions and an anonymous referee for his careful reading and insigthful comments to improve the manuscript.

# References


[1] C. Abraham, G. Biau, and B. Cadre. On the kernel rule for function classification. *Annals of the Institute of Statistical Mathematics*, 58:619–633, 2006.

[2] J. Bardaji. Etude de la méthode des forêts aléatoires à des fins de prévision de la production manufacturière. In *N° 036/DG75-G120*, Insee, February 2007.

[3] G. Biau, L. Devroye, and G. Lugosi. Consistency of random forests and other averaging classifiers. 2007. Preprint, University Paris VI.

[4] O. Biau, H. Erkel-Rousse, and N. Ferrari. Individual responses to business tendency surveys and the forecasting of manufactured production: An assessment of the Mitchell, Smith and Weale dis-aggregate indicators on French data. *Economie et Statistique*, 395-396:91–116, 2006.

[5] L. Breiman. Statistical modeling: The two cultures. *Statistical Science*, 13:119–215, 2001a.

[6] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001b.

[7] L. Breiman, J.H. Friedman, R.A. Olsen, and C.J. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, 1984.

[8] S. Buffeteau and V. Mora. Predicting the national accounts of the euro zone using business surveys. *Conjoncture in France, INSEE*, December 2002.

[9] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New-York, 1996.

[10] E. Dubois and E. Michaux. Etalonnages à l'aide d'enquêtes de conjoncture : de nouveaux résultats. *Economie et Prévision*, 172:11–28, 2006.

[11] B. Efron and R.J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, New York, 1993.

[12] T. Hastie, R.J. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

[13] Y. Lin and Y. Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101:578–590, 2006.

[14] J. Mitchell, R.J. Smith, and M.R. Weale. Aggregate versus disaggregate survey-based indicators of economic activity. In *27th CIRET conference*, Warsaw, September 2004.

[15] J. Mitchell, R.J. Smith, and M.R. Weale. Forecasting manufacturing output growth using firm-level survey data. *The Manchester School*, 73(4):479–499, 2005.

[16] M. Reynaud and S. Scherrer. Une modélisation VAR de l'enquête de conjoncture de l'INSEE dans l'industrie. *Document de travail de la Direction de la Prévision*, 96-12, 1996.